BERNSTEIN
SOCIETE GENERALE GROUP

27 April 2026

China Internet

China Internet: AI inference margins addendum; DeepSeek-V4, hy3-preview thoughts

Robin Zhu
+852 2123 2659
robin.zhu@bernsteinsg.com

Charles Gou
+852 2123 2618
charles.gou@bernsteinsg.com

Min-Joo Kang
+852 2123 2644
minjoo.kang@bernsteinsg.com

Feedback on our AI token economics model. Our recent AI token economics primer (LINK) drove considerable discussion and debate with investors, as well as feedback from some of the top Chinese AI labs. While our first effort was mostly intended as a bare-bones estimate aimed at demonstrating how the math behind inference margins worked, the feedback has offered useful real-world insight on where a few hard-to-measure assumptions land in practice.

Refreshing our tokens per second throughput assumptions. The biggest change in this note versus our initial H1 2026 estimates relate to higher tokens-per-second (TPS) assumptions across the board - reflecting direct company feedback on datacentre level throughput. Qualitatively, our updated assumptions imply a small inference margin delta between Z.ai and Minimax than previously estimated. In contrast with the 3x ratio we'd originally assumed between M2.5/2.7 and Z.ai GLM-5/5.1 TPS, and a 4x ratio in active parameters, reconciling guidance for 40% text/coding margins for M2.7 requires a 4.5-5.0x ratio in relative datacentre level TPS ratio between M2.7 and GLM5/5.1. We've added third-party throughput data and sensitivity analysis in this note.

Commercial discounts, and the conclusions that don't change. We intentionally ignored commercial discounts in our previous modeling, as they're not directly tied to how AI models consume and generate tokens, and somewhat unknowable outside-in. But it's clear that customer mix and discounting significantly impacts revenue per token. Compared with our previous effort, we think our more important takeaways survive - that the GLM family of models likely generates higher margins assuming similar discounting, and a relative TPS ratio inversely correlated to active parameters per token. Multi-modal inference generates much higher rev/Mtok and inference margins.

Godot arrives - some early thoughts on DeepSeek V4. Our first glance take on DeepSeek-V4 boils down to "close to global SOTA from a few months ago, at significantly lower cost". But - as has been foreshadowed for a while - the degree of separation between DeepSeek and domestic peers (e.g. Qwen, Z.ai, Kimi) has narrowed. V4-Flash priced at the lower end of the peer range, which reinforces our bias that price competition among lighter, low-cost models will remain fierce. V4-Pro pricing was initially adjacent to domestic peers like Z.ai and Kimi, but a 75% "time limited" price cut immediately after launch felt noteworthy too.

Tencent starts climbing the curve. We think the best framing of hy3-preview is "a somewhat belated step in the right direction". Our sense is the results were fine relative to modest expectations - if not necessarily delivering the kind of "wow factor" Vinces Yao's pedigree had led some investors to anticipate. The real proof in the pudding will be whether Tencent's rebuild of its AI data and infra organization can enable faster iteration of in-house model capabilities… and when the company might be able to deliver on its own "Muse Spark moment" in the coming months. Our bias remains that reprogramming the Internet top funnel in China will take longer.

First Published: 26 Apr 2026 20:30 UTC Completion Date: 26 Apr 2026 19:21 UTC

Robin Zhu +852 2123 2659 robin.zhu@bernsteinsg.com 27 April 2026

BERNSTEIN TICKER TABLE

			24 Apr 2026		TTM	Adjusted EPS				Adjusted P/E (x)	
			Closing	Price	Rel.						
Ticker	Rating	Cur	Price	Target	Perf.	Cur	2025A	2026E	2027E	2025A	2026E	2027E
700.HK (Tencent)	O	HKD	493.40	780.00	(39.3)%	CNY	28.09	29.89	34.56	15.3	14.4	12.5
BABA (Alibaba)	O	USD	135.82	180.00	(17.5)%	CNY	65.41	28.97	42.38	14.2	32.0	21.9
9988.HK (Alibaba)	O	HKD	131.80	176.00	(27.5)%	HKD	8.40	3.91	5.71	13.7	29.4	20.1
ASIAX			1,827.77									
SPX			7,108.40									

O - Outperform, M - Market-Perform, U - Underperform, NR - Not Rated, CS - Coverage Suspended

Source: Bloomberg, Bernstein estimates and analysis.

INVESTMENT IMPLICATIONS

In this note we've updated our AI token economics framework to incorporate feedback from China's AI labs. Qualitatively, the 40% gross margin assumption Minimax has guided implies a materially higher tokens-per-second throughput than we previously assumed. While this can be explained in a number of ways, centred around hardware and software optimization, our bias remains that Z.ai's frontier models should have higher margins, assuming relative TPS ratios banded around active parameters per token scale, typical workload mix, and comparable commercial discount levels. The shift from multi-modal to text/coding workloads

Looking top down, the DeepSeek V4 release keeps China within months of the global SOTA, while the open source, open weights nature of its release should help to elevate capabilities across the leading Chinese AI labs. Tencent's hy3-preview release took place amid very different levels of ex ante expectations, and should probably be considered through the lens of whether it signals faster model and application layer iteration in the coming months.

VALUATION COMPS TABLE

EXHIBIT 1: **Asia Internet: valuation summary**

	Rating	Price target	Last price	Crncy	Market cap (US$mn)	PE 2026E	PE 2027E	PE 2028E	EV/sales 2026E	EV/sales 2027E	EV/sales 2028E
China Internet coverage											
Tencent	O	780	493.40	HKD	574,740	14.6x	12.6x	10.9x	4.8x	4.2x	3.7x
PDD	M	132	98.03	USD	139,168	8.2x	7.4x	6.6x	0.9x	0.6x	0.4x
Meituan	M	85	82.45	HKD	64,982	n.a.	23.5x	12.3x	1.0x	0.9x	0.7x
NetEase	O	150	110.58	USD	70,564	12.8x	11.4x	10.6x	2.9x	2.6x	2.4x
Boss Zhipin	M	16.5	13.62	USD	6,585	11.0x	9.0x	7.9x	2.5x	1.8x	1.3x
JD	O	36	30.27	USD	41,343	10.3x	7.7x	6.0x	0.2x	0.2x	0.2x
Alibaba	O	180	135.82	USD	325,839	31.4x	22.0x	19.4x	2.2x	2.0x	1.9x
China Internet other											
Kuaishou			43.70	HKD	24,249	9.7x	8.5x	7.6x	1.0x	0.9x	0.9x
Bilibili			22.25	USD	9,457	23.4x	17.6x	13.3x	1.5x	1.4x	1.3x
TME			9.34	USD	14,700	9.8x	8.6x	7.7x	1.9x	1.8x	1.6x
iQiyi			1.17	USD	1,129	57.5x	10.2x	5.4x	0.7x	0.6x	0.6x
Baidu			128.71	USD	43,794	16.2x	14.0x	11.5x	1.4x	1.3x	1.2x
VIPshop			14.36	USD	6,897	5.6x	5.3x	5.1x	0.3x	0.3x	0.3x
Tencent Music			9.34	USD	14,700	9.8x	8.6x	7.7x	1.9x	1.8x	1.6x
Trip.com			53.11	USD	34,713	13.2x	11.6x	10.2x	2.7x	2.4x	2.1x
KE Holdings			16.18	USD	18,887	21.4x	18.1x	17.0x	1.1x	1.1x	1.0x
Asian Internet											
Naver	O	300,000	215,000	KRW	22,843	14.9x	13.3x	12.8x	1.9x	1.7x	1.4x
Kakao	O	80,000	48,200	KRW	14,464	31.7x	26.0x	19.9x	1.8x	1.5x	1.2x
Hybe	O	500,000	255,000	KRW	7,445	21.3x	21.4x	17.1x	2.5x	2.1x	1.9x
Coupang	U	17	20.51	USD	37,496	123.3x	44.8x	35.2x	0.9x	0.8x	0.7x
Sea Ltd.			85.44	USD	52,331	23.9x	17.9x	14.5x	1.5x	1.3x	1.1x
Grab			3.90	USD	15,992	38.6x	24.7x	18.1x	2.8x	2.3x	1.9x
US Internet											
Amazon			263.99	USD	2,838,219	29.3x	24.3x	20.3x	3.6x	3.2x	3.0x
Alphabet			344.40	USD	4,160,456	28.2x	24.4x	20.9x	10.1x	8.6x	7.4x
Meta			675.03	USD	1,713,512	20.4x	17.4x	14.1x	6.9x	5.8x	5.0x
Netflix			92.44	USD	389,246	26.0x	24.0x	20.2x	7.7x	6.9x	6.2x
Uber			74.64	USD	152,029	21.9x	16.9x	13.4x	2.7x	2.4x	2.1x
Spotify			518.00	USD	106,621	33.4x	27.3x	22.8x	4.3x	3.8x	3.4x
DoorDash			176.78	USD	77,031	32.5x	25.0x	21.2x	4.2x	3.5x	3.0x

Pricing date April 22, 2026. The valuation multiples of our China and Asian Internet coverage are based on Bernstein estimates; the other companies shown reflect Bloomberg consensus estimates.
Source: Corporate reports, Bloomberg, Bernstein estimates and analysis.

DETAILS

ITERATING OUR AI TOKEN ECONOMICS AND INFERENCE MARGIN MODELING

For a while now we've compared following frontier AI developments to the "blind men and the elephant" parable, while the speed of day-to-day news flow has remained rapid. In this context, we've been grateful for the feedback from contacts at China's top AI labs on the token economics framework we recently published (LINK). In the main, the feedback has reinforced our confidence in the mechanics of the framework, while offering useful real-world insight on how a few hard-to-measure assumptions work in practice.

Rationalising Minimax's guidance for 40% text/coding inference margins on M2.7

Most notably, we've updated our H1 2026 projections to reflect feedback from Minimax, where management pointed to 40% inference margins for M2.7 (see Exhibit 2). Our discussions with the company reinforced assumptions we'd made like GPU cost per hour ($1.6 per GPU per hour, assuming a H100/H800 stack), but the company argued that a combination of M2.5/2.7's smaller parameter scale and in-house optimization work (focused on memory and cache use, and custom hardware infrastructure) meant TPS was materially higher than we'd assumed.

In hindsight, we'd likely underestimated *both* Z.ai and Minimax TPS at the datacentre level. Techniques like quantization (model compression) and prefill-decode separation (enabling better hardware optimisation, and batching/scheduling at the model level) are use commonly across the top AI labs to improve token throughput. On a relative basis, a 4x ratio between M2.5/2.7 and GLM-5/5.1 TPS - to match the ratio between active parameters per token scale - would probably have been a better starting assumption. The 4.5-5.0x ratio in relative TPS that Minimax argued for - and is required for M2.7 to achieve 40% inference margins using our framework - is higher, but ultimately remains in the realm of plausible, based on third-party data.

To illustrate, benchmarking from AI infrastructure provider Lambda Labs pointed to TPS of 8,062 tok/s for M2.5, assuming a pair of Nvidia B200 GPUs. That compares with 6,300 tok/s for GLM-5 on an Nvidia HGX B200 stack which incorporates eight B200 GPUs. Dividing both by the respective number of GPUs points to 4,031 tok/s per GPU for M2.5, roughly 5.1 times 788 tok/s for GLM-5 (see Exhibit 3 and Exhibit 4). The query volumes that Z.ai and Minimax cater to at the datacentre level are almost certainly *both* much higher than the Lambda.ai benchmark, which directionally argues for TPS levels that are higher for both labs.

Addressing the influence of commercial discounts

We intentionally left commercial discounts out of our initial framework for the sake of simplicity. But as we move closer to modeling at the individual AI lab level, the inclusion of a commercial discounting variable feels apt - at least as something our readers can flex. Assuming higher discounts obviously depresses inference margins. Directionally it's probably reasonable to assume higher average discounting on large enterprise customer workloads than small and medium developers buying coding plans (which are still generally sold out as more supply is made available). But in practice when analysing the AI labs time-limited discounts aimed at user acquisition are probably best considered dynamically. Exhibit 5 and Exhibit 6 outline our updated token economics equation and an explanation of the various variables. We've added a commercial discount variable to our model - we've assumed a flat 20% across the board as a base case, and show sensitivity analysis later in the note, reflecting different levels of discounting. We've also renamed TPS (previously "T") to align with our discussions with the AI labs.

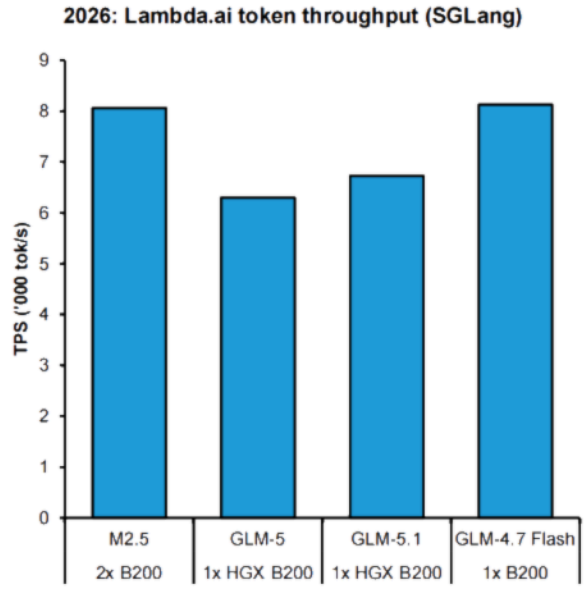
Several of the discussions we've had with industry contacts have pointed to the fluidity of compute costs booked in cost of sales versus R&D, and discounting versus marketing expenses. Other variables we were reminded of that influence all-in inference margins include maintenance costs and other corporate level overheads - we've continued to exclude these from our model for simplicity's sake. All in, we'd still urge investors to consider our estimates indicative rather than precise financial projections. But compared with our previous effort, these updated conclusions should be closer aligned with the combination of company feedback and observable data points in the field.

EXHIBIT 2: **We've updated our AI token economics modelling to reflect feedback from the top Chinese AI labs, and introduced a new commercial discount variable**

		Z.ai				Minimax	
Variable		GLM-4.5	GLM-4.7	GLM-5	GLM-5.1	M2.5	M2.7
Rev/Mtok	Revenue per million tokens	$0.42	$0.42	$0.67	$0.93	$0.13	$0.15
IO	Input vs. output sequence length ratio	15	15	15	15	20	20
P-in	Input token list price per million tokens	$0.60	$0.60	$1.00	$1.40	$0.30	$0.30
P-out	Output token list price per million tokens	$2.20	$2.20	$3.20	$4.40	$1.20	$1.20
KV	KV cache hit rate	40%	40%	40%	40%	70%	70%
D-kv	Cache read vs. input token price	20%	20%	20%	20%	10%	20%
C	Commercial discount	20.0%	20.0%	20.0%	20.0%	20.0%	20.0%
Cost/Mtok	Inference cost per million tokens	$0.35	$0.35	$0.31	$0.28	$0.07	$0.05
TPS	Tokens per second per GPU	1,600	1,600	1,800	2,000	7,500	9,000
U	Utilisation	80%	80%	80%	80%	90%	90%
Cost-GPU-hr	GPU cost per hour	1.6	1.6	1.6	1.6	1.6	1.6
Margin		-3.5%	-3.5%	33.9%	50.3%	29.5%	42.5%

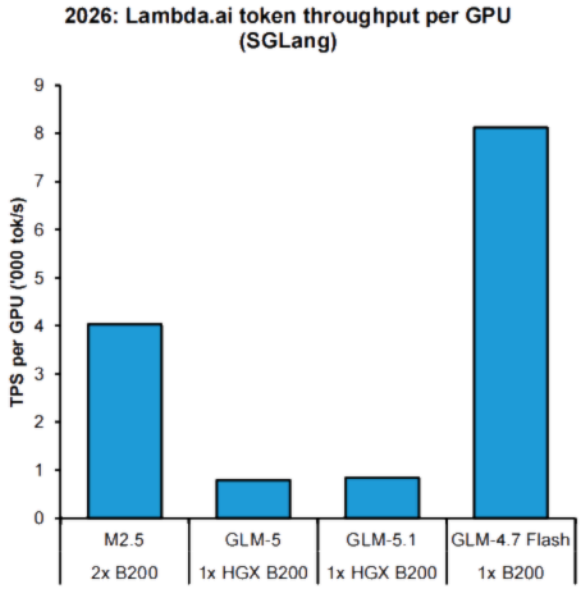
Source: Bernstein analysis.

EXHIBIT 3: **Lambda.ai interestingly shows benchmarking data for deployment-level throughput**



Source: Lambda.ai and Bernstein analysis.

EXHIBIT 4: **This shows M2.5 TPS roughly five times that of GLM-5, while Z.ai's flash models are faster**



Source: Lambda.ai and Bernstein analysis.

EXHIBIT 5: **We've left our assumptions for commercial discounting at zero for now, but expect this to be a useful plug when analysing reported AI lab financials going forward**

$$Rev/Mtok = \left[\underbrace{\frac{IO}{(IO+1)}}_{\text{Input token mix}} \times P_{in} \times \underbrace{\left((1-KV) + KV \times D_{kv}\right)}_{\text{Discount for KV cache tokens}} + \underbrace{\frac{1}{(IO+1)}}_{\text{Output token mix}} \times P_{out}\right]_{\text{Input token cost}} \times (1-C)$$

$$Cost/Mtok = \frac{Cost_{GPU-hr}}{\underbrace{(TPS \times U \times 3600/10^6)}_{\text{Tokens per GPU-hour}}}$$

Source: Bernstein analysis.

EXHIBIT 6: **The most contentious assumption in our previous analysis was tokens per second throughput, which can reflect a variety of hardware and software optimisation vectors**

Variable		Comments
Rev/Mtok	**Revenue per million tokens**	
IO	**Input vs. output sequence length ratio**	• Ratio between input and output token lengths • Typically higher for coding and agentic tasks, shorter for general reasoning tasks with longer reasoning traces and end output
P-in	**Input token price per million tokens**	• Listed input token price less commercial discounts
P-out	**Output token price per million tokens**	• Listed output token price less commercial discounts
KV	**KV cache hit rate**	• Listed cache read token price less commercial discounts • KV cache hit ratios can vary considerably by workload, typically higher for repetitive agentic workloads, lower for general reasoning
D-kv	**Cache read token discount vs. input token price**	• Listed output token price less commercial discounts, expressed as a percentage of non-cached input token price
C	**Commercial discount**	• Reflects commercial strategy, timed promotions, customer mix, etc.
Cost/Mtok	**Inference cost per million tokens**	
T	**Tokens per second per GPU**	• Tokens calculated per second per GPU • Can be influenced by model architecture (e.g. parameter scale), batching, etc.
U	**Utilisation**	• Assumed average GPU utilisation rate
Cost-GPU-hr	**GPU cost per hour**	• Cost of GPU compute rental per hour, inclusive of overhead costs and hyperscaler margin if applicable

Source: Bernstein analysis.

SENSITIVITY ANALYSIS, AND OUR CONCLUSIONS THAT (WE THINK) DON'T CHANGE

In Exhibit 7 to Exhibit 10 we've shown sensitivity analyses of the implied GLM-5, GLM-5.1, M2.5, and M2.7 margins, assuming different levels of TPS and blended commercial discounts. Compared with our previous note, we think our more important takeaways survive. We still think the GLM family of models likely generates higher margins assuming comparable level of discounting, and relative TPS ratios within a band around the inverse of active parameters per token scale. Minimax commentary pointed to higher multi-modal than text/coding inference margins, which was consistent with our view that video generation and text-to-speech generate higher rev/Mtok (e.g. since GPUs costs are the same). The company also confirmed that text/coding workloads have driven a larger share of growth year to date, driven by the surge in interest in OpenClaw and broader agentic AI this year (see Exhibit 11).

Our $1.6 estimate for GPU cost per hour (which we arrived at through triangulation) was, encouragingly, generally considered sensible. Our preference here remains to assume largely similar costs across the top Chinese AI labs, given they tend to have

similar access to suppliers of GPU compute. We expect comparisons of reported versus theoretical inference margins going forward to yield interesting insights on the extent to which different AI labs can drive software and hardware level optimization.

EXHIBIT 7: **The way we've rationalised 40% M2.7 gross margins has been via much higher assumptions for token throughput, supported in part by the Lambda benchmark results**

Minimax M2.7

Tokens /sec \ Blended commercial discount (%)	10.0%	12.5%	15.0%	17.5%	20.0%	22.5%	25.0%	27.5%	30.0%
7,000	47.1%	43.4%	39.6%	35.7%	31.8%	27.7%	23.6%	19.3%	14.9%
7,500	50.0%	46.3%	42.6%	38.9%	35.0%	31.0%	27.0%	22.8%	18.6%
8,000	52.5%	48.9%	45.3%	41.6%	37.8%	33.9%	30.0%	25.9%	21.8%
8,500	54.7%	51.2%	47.6%	44.0%	40.3%	36.5%	32.6%	28.7%	24.6%
9,000	56.7%	53.2%	49.7%	46.1%	42.5%	38.8%	35.0%	31.1%	27.1%
9,500	58.4%	55.0%	51.6%	48.0%	44.5%	40.8%	37.1%	33.3%	29.4%
10,000	60.0%	56.6%	53.2%	49.8%	46.2%	42.7%	39.0%	35.3%	31.4%
10,500	61.4%	58.1%	54.7%	51.3%	47.8%	44.3%	40.7%	37.0%	33.3%
11,000	62.7%	59.4%	56.1%	52.7%	49.3%	45.8%	42.3%	38.6%	34.9%

Source: Corporate reports, Bernstein estimates and analysis.

EXHIBIT 8: **Similarly, we've assumed a higher TPS for Minimax's M2.5 model in H1 2026, though implied inference margins can vary meaningfully on different TPS and discounting assumptions**

Minimax M2.5

Tokens /sec \ Blended commercial discount (%)	10.0%	12.5%	15.0%	17.5%	20.0%	22.5%	25.0%	27.5%	30.0%
5,500	28.7%	24.5%	20.1%	15.7%	11.1%	6.4%	1.5%	-3.5%	-8.8%
6,000	33.8%	29.7%	25.5%	21.2%	16.8%	12.3%	7.6%	2.8%	-2.2%
6,500	38.2%	34.2%	30.1%	26.0%	21.7%	17.3%	12.8%	8.2%	3.4%
7,000	41.9%	38.0%	34.0%	30.0%	25.9%	21.6%	17.2%	12.8%	8.1%
7,500	45.1%	41.3%	37.4%	33.5%	29.5%	25.3%	21.1%	16.7%	12.2%
8,000	47.9%	44.2%	40.4%	36.6%	32.6%	28.6%	24.5%	20.2%	15.9%
8,500	50.4%	46.7%	43.0%	39.3%	35.4%	31.5%	27.4%	23.3%	19.0%
9,000	52.6%	49.0%	45.4%	41.7%	37.9%	34.0%	30.1%	26.0%	21.9%
9,500	54.5%	51.0%	47.4%	43.8%	40.1%	36.3%	32.4%	28.5%	24.4%

Source: Corporate reports, Bernstein estimates and analysis.

EXHIBIT 9: **We still think Z.ai likely has higher margins than Minimax's models, assuming comparable levels of commercial discounting, though the relative gap is likely smaller than we'd previously estimated**

Z.ai GLM-5

Tokens /sec \ Blended commercial discount (%)	10.0%	12.5%	15.0%	17.5%	20.0%	22.5%	25.0%	27.5%	30.0%
1,400	37.4%	33.3%	29.3%	25.1%	20.8%	16.4%	11.8%	7.1%	2.3%
1,500	40.9%	37.0%	33.0%	28.9%	24.7%	20.4%	16.0%	11.5%	6.8%
1,600	43.9%	40.1%	36.2%	32.2%	28.2%	24.0%	19.7%	15.3%	10.8%
1,700	46.6%	42.9%	39.1%	35.2%	31.2%	27.2%	23.0%	18.7%	14.3%
1,800	49.1%	45.4%	41.6%	37.8%	33.9%	29.9%	25.9%	21.7%	17.4%
1,900	51.2%	47.6%	43.9%	40.2%	36.4%	32.5%	28.4%	24.3%	20.1%
2,000	53.1%	49.6%	46.0%	42.3%	38.5%	34.7%	30.8%	26.8%	22.6%
2,100	54.9%	51.4%	47.8%	44.2%	40.5%	36.7%	32.9%	28.9%	24.9%
2,200	56.5%	53.0%	49.5%	46.0%	42.3%	38.6%	34.8%	30.9%	26.9%

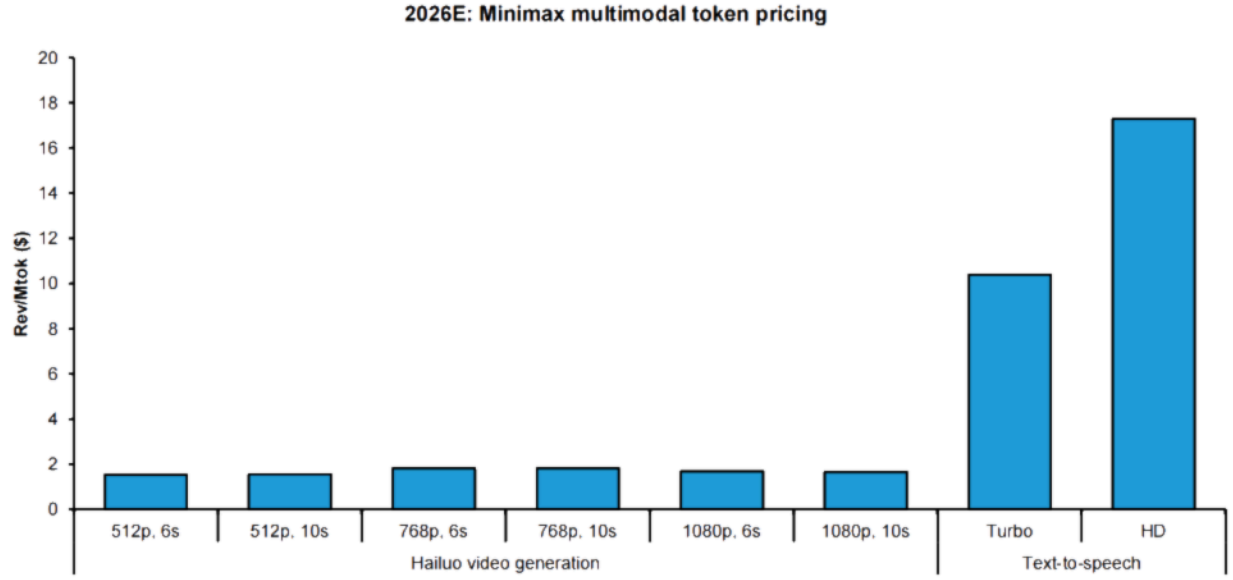
Source: Corporate reports, Bernstein estimates and analysis.

EXHIBIT 10: **Assuming M2.7 has higher inference margins would require it to have an even greater TPS gap to Z.ai's models than we've now assumed**

	Z.ai GLM-5.1 Blended commercial discount (%)								
Tokens /sec	10.0%	12.5%	15.0%	17.5%	20.0%	22.5%	25.0%	27.5%	30.0%
1,600	57.0%	53.5%	50.0%	46.5%	42.8%	39.1%	35.3%	31.5%	27.5%
1,700	58.9%	55.5%	52.1%	48.6%	45.0%	41.4%	37.7%	33.9%	30.0%
1,800	60.6%	57.3%	53.9%	50.5%	47.0%	43.4%	39.8%	36.0%	32.2%
1,900	62.2%	58.9%	55.5%	52.1%	48.7%	45.2%	41.6%	38.0%	34.2%
2,000	63.6%	60.3%	57.0%	53.7%	50.3%	46.8%	43.3%	39.7%	36.0%
2,100	64.8%	61.6%	58.3%	55.0%	51.7%	48.3%	44.8%	41.2%	37.6%
2,200	66.0%	62.8%	59.6%	56.3%	53.0%	49.6%	46.2%	42.7%	39.1%
2,300	67.0%	63.9%	60.7%	57.4%	54.1%	50.8%	47.4%	44.0%	40.4%
2,400	68.0%	64.8%	61.7%	58.5%	55.2%	51.9%	48.6%	45.2%	41.7%

Source: Corporate reports, Bernstein estimates and analysis.

EXHIBIT 11: **Minimax confirm to us that its multi-modal workloads generate higher gross margins than text/ coding... and that the latter has driven a larger share of incremental growth in 2026 to date**



Source: Corporate reports and Bernstein analysis.

OBSERVATIONS ON DEEPSEEK V4 AND HY3-PREVIEW

Exhibit 12 and Exhibit 13 show comparisons of token pricing across the top AI labs, contrasting the DeepSeek V4-Pro and V4-Flash models, and hy3-preview, with the spread of domestic peers. At a high level, the DeepSeek V4 pricing at release reinforces our biases on the competitive outlooks for frontier intelligence (where we think some degree of pricing power will persist for the leaders) versus the lighter, low-cost backbone end of the market, where fierce price competition feels a given. Strategically, we continue to favour frontier intelligence as the most important arbiter of long-term success in this space.

DeepSeek V4-Pro pricing adjusted for "typical" workload mix looks largely comparable with other leading Chinese models like GLM-5.1 and Qwen3.6 Max, at around $1 rev/Mtok. The fact a "time limited" 75% discount was introduced on domestic (but not international pricing) less than 24 hours after launch was telling though, as far as user acquisition strategy goes. DeepSeek V4-Flash was priced near the lower bound of lighter, faster models including Minimax and various Flash models. Tencent's hy3-preview model is probably best interpreted as an interim product on the path to larger models, but has also priced towards the lower end of the peer range.

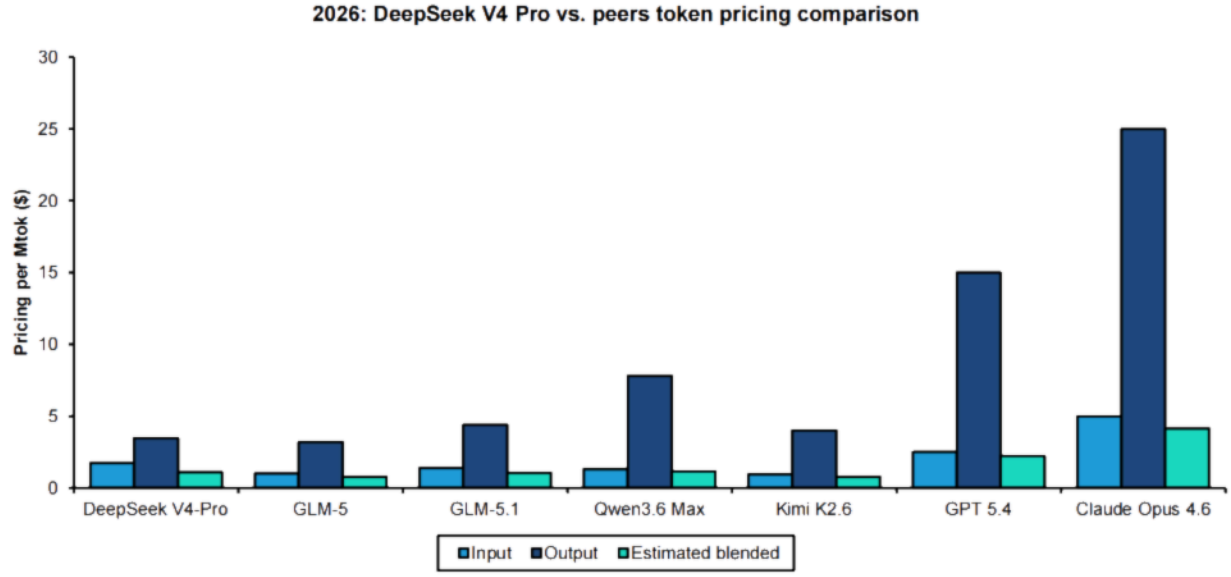
Thoughts on the "now what" question

The more interesting observations we've heard on DeepSeek V4 have included the absence of any claims it was trained on domestic silicon, and DeepSeek's claim that "DeepSeek-V4-Pro requires 27% of single-token inference FLOPs and 10% of

the KV cache compared with DeepSeek-V3.2". The pathways DeepSeek found to achieve this are beyond the scope of this note, but deserves attention given the importance of long-term memory to long-duration task completion. For the broader sector, while aggressive DeepSeek pricing obviously represents a source of competitive pressure, the fact the model was presented on an open source/weights basis should help to accelerate progress across the broader AI frontier. Our base case for DeepSeek's upcoming fundraising is that it will be led by Tencent, given the working relationship between the two companies (e.g. in Yuanbao), and Alibaba's investments in Kimi and Minimax.

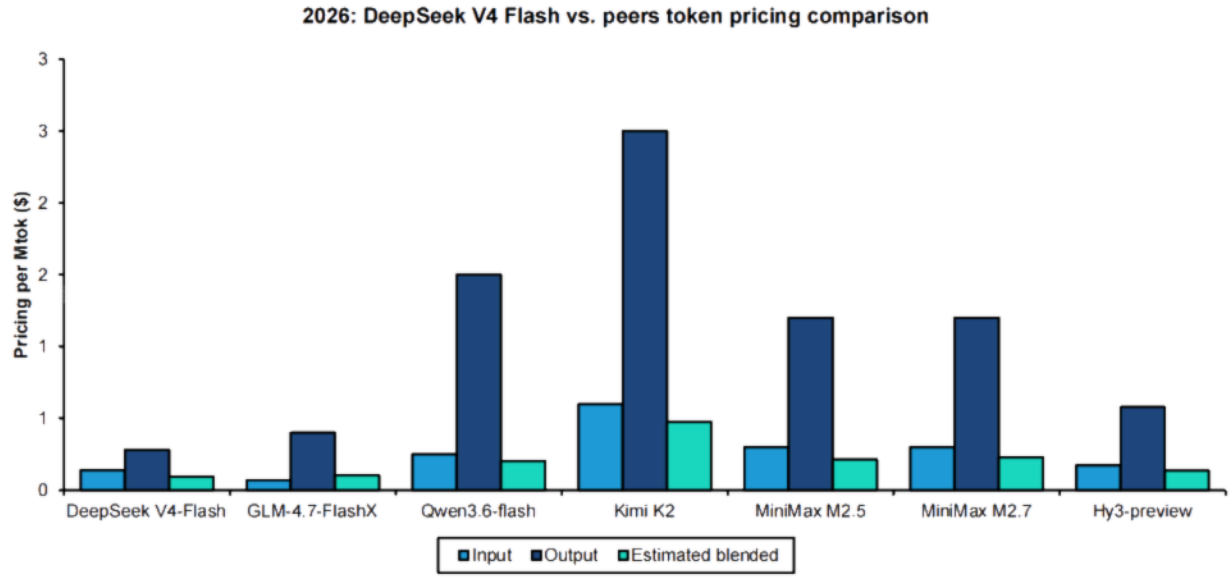
Beyond benchmark scores and token pricing, we think the test of the value of hy3-preview and its subsequent full release will be the extent to which Tencent's rebuild of its AI data and infra organization can now enable faster iteration of in-house model capabilities… and when Tencent might be able to deliver on its own "Muse Spark moment" a larger (1T?) model. The company's development of agentic functionality in WeChat has long been somewhat model agnostic… but remains another area where the lack of new news has been a narrative headwind. Meanwhile, we've continued to hear positive feedback on the company's ability to utilise AI in its ads business. The fact investors largely ignored Tencent's release of its HY-World 2.0 model (which creates in-game assets that can be exported to Unreal and Unity) was unsurprising, but felt like another example of how narrow investor focus on AI has become.

EXHIBIT 12: **DeepSeek V4-Pro has priced in line with peers like GLM-5 and Qwen3.6 Max, with blended rev/Mtok slightly over $1**



Source: Corporate reports and Bernstein analysis.

EXHIBIT 13: **DeepSeek V4-Flash (and hy3-preview) pricing look meaningfuly lower than existing incumbents, which feels aligned with our view that the market for lighter, low-cost compute will become increasingly price competitive**



Source: Corporate reports and Bernstein analysis.